

General and Robust Communication-Efficient Algorithms for Distributed Clustering *

Pranjal Awasthi

Maria-Florina Balcan

Colin White

Abstract

As datasets become larger and more distributed, algorithms for distributed clustering have become more and more important. In this work, we present a general framework for designing distributed clustering algorithms that are robust to outliers. Using our framework, we give a distributed approximation algorithm for k -means, k -median, or generally any ℓ_p objective, with z outliers and/or balance constraints, using $O(m(k+z)(d+\log n))$ bits of communication, where m is the number of machines, n is the size of the point set, and d is the dimension. This generalizes and improves over the previous work of Bateni et al. [12] and Malkomes et al. [31]. As a special case, we achieve the first distributed algorithm for k -median with outliers, answering an open question posed by Malkomes et al. [31]. For distributed k -means clustering, we provide the first dimension-dependent communication complexity lower bound for finding the optimal clustering. This improves over the lower bound of Chen et al. which is dimension-agnostic [18].

Furthermore, we give distributed clustering algorithms which return nearly optimal solutions, provided the data satisfies the approximation stability condition of Balcan et al. [8] or the spectral stability condition of Kumar and Kannan [27]. In certain clustering applications where each machine only needs to find a clustering consistent with the global optimum, we show that no communication is necessary if the data satisfies approximation stability.

*Authors' addresses: pranjal.awasthi@rutgers.edu, ninamf@cs.cmu.edu, crwhite@cs.cmu.edu. This work was supported in part by NSF grants CCF-1422910, CCF-1535967, IIS-1618714, a Sloan Research Fellowship, a Microsoft Research Faculty Fellowship, a Google Research Award, and a National Defense Science & Engineering Graduate (NDSEG) fellowship.

1 Introduction

Clustering is a fundamental problem in machine learning with applications in many areas including computer vision, text analysis, bioinformatics, and so on. The underlying goal is to group a given set of points to maximize similarity inside a group and dissimilarity among groups. A common approach to clustering is to set up an objective function and then approximately find the optimal solution according to the objective. Examples of these objective functions include k -means, k -median, and k -center, and more generally any ℓ_p objective, in which the goal is to find k centers to minimize the sum of the ℓ_p distances from each point to its closest center. Motivated by real-world constraints, further variants of clustering have been studied. For instance, in k -clustering with outliers, the goal is to find the best clustering (according to one of the above objectives) after removing a specified number of datapoints, which is useful for noisy data or outliers. The capacitated clustering variant adds the constraint that no individual cluster can be larger than a certain size, which is useful for load-balancing. Finding approximation algorithms to different clustering objectives and variants has attracted significant attention in the computer science community [3, 15, 16, 17, 19, 23, 30].

As datasets become larger, sequential algorithms designed to run on a single machine are no longer feasible for real-world applications. Additionally, in many cases data is naturally spread out among multiple locations. For example, hospitals may keep records of their patients locally, but may want to cluster the entire spread of patients across all hospitals in order to do better data analysis and inference. Therefore, distributed clustering algorithms have gained popularity over the past few years [11, 12, 31]. In the distributed setting, it is assumed that the data is partitioned arbitrarily across m machines, and the goal is to find a clustering which approximates the optimal solution over the entire dataset while minimizing communication among machines. Recent work in the theoretical machine learning community establishes guarantees on the clusterings produced in distributed settings for certain problems [11, 12, 31]. For example, Malkomes et al. provide distributed algorithms for k -center and k -center with outliers [31], and Bateni et al. introduce distributed algorithms for capacitated k -clustering under any ℓ_p objective [12]. A key algorithmic idea common among both of these works is the following: each machine locally constructs an approximate size $\tilde{O}(k)$ summary of its data; the summaries are collected on a central machine which then runs a sequential clustering algorithm. A natural question that arises is whether there is a unifying theory for all distributed clustering variants. In the current work, we answer this question by providing a general distributed algorithm for clustering under any ℓ_p objective with or without outliers, and with or without capacity constraints, thereby generalizing and improving over recent results. We complement our algorithm by giving the first dimension-dependent lower bound on the level of communication needed to find the optimal distributed clustering, improving over the recent dimension-agnostic lower bound of Chen et al. [18].

Certain real-world applications might require nearly optimal clusterings, closer than the constant-factor worst-case approximation ratios mentioned in the previous paragraph. However, these approximation ratios are unavoidable in the worst case due to existing lower bounds, for example, k -center cannot be approximated to a factor smaller than $2 - \epsilon$ even on a single machine [23]. To go beyond these worst-case results, many recent works have studied natural structure that exists in real-world instances, and shown that algorithms can output a clustering very close to optimal under these natural stability assumptions [5, 7, 8, 10, 13, 21, 27, 37]. For example, the (c, ϵ) -approximation stability condition defined by Balcan et al. states that any c -approximation to the clustering objective is ϵ -close to the target clustering [1, 8, 25], and the spectral stability condition of Kumar and Kannan is a deterministic generalization of many generative clustering models [27, 34, 35]. In the current work, we study these conditions in the distributed setting, and design algorithms with low communication that have guarantees just as strong as in the sequential setting. Finally, motivated

by several examples, we initiate the study of a relaxed version of distributed clustering in which each machine needs to find a clustering consistent with the global optimal, which we call a *locally consistent clustering*. Surprisingly, we show that if the data satisfies approximation stability, *no* communication is necessary for each machine to find a locally consistent clustering.

1.1 Our contributions

In this paper, we provide conceptually simple distributed clustering algorithms, combined with a new analysis, to paint a unifying picture for distributed clustering that generalizes and improves over several recent results. A distributed clustering instance consists of a set of n points partitioned arbitrarily across m machines, with a distance metric (if points lie in \mathbb{R}^d , the metric is Euclidean distance) and the problem is to approximate the ℓ_p clustering objective while minimizing the amount of communication across machines. For clustering with capacities or outliers, the number of outliers z and the capacity constraint L are inputs to the algorithm (as is k).

General Robust Distributed Clustering. In Section 3, we show a general distributed algorithm for balanced k -clustering in ℓ_p in d dimensions with z outliers, using $O(m(k+z)(d+\log n))$ communication. The algorithm can be summarized as follows. Each machine performs a k -clustering on its own data, and then sends the centers, along with the sizes of their corresponding clusters, to a central machine. The central machine then runs a weighted clustering algorithm on the mk centers. We add necessary changes to handle capacities and outliers, for instance, each (non-central) machine runs a $k+z$ clustering algorithm in the case of k -clustering with z outliers. Given a sequential α -approximation algorithm and a bicriteria β -approximation algorithm which opens up $\tilde{O}(k)$ centers, in Theorem 3 we prove our distributed algorithm returns an $O(\alpha\beta)$ approximation and improves over the approximation guarantees of Bateni et al. [12]. For example, for k -median, we achieve a $(6\alpha + 2 + \epsilon)$ -approximation by plugging in the bicriteria algorithm of Lin and Vitter [29] (adding an $O(\log n)$ factor to the communication cost), as opposed to a 32α -approximation from Bateni et al. [12]. By plugging in the approximation algorithm for k -median with outliers [19], our algorithm achieves the first constant approximation for distributed k -median with outliers, answering an open question posed by Malkomes et al. [31]. We achieve these improvements by using a refined analysis that applies the triangle inequality in clever ways to prove a strong bound on the distance between each local center and its closest center in the solution outputted by the algorithm. We show how to carefully reason about the optimal clustering with subsets of outliers removed to preserve the constant factor approximation guarantee for k -median with outliers.

Communication Complexity Lower Bounds. In Section 4, we consider the communication complexity of distributed clustering – the minimum number of bits that need to be communicated among the machines by any algorithm which can solve the problem on all inputs with high probability. Recently, Chen et al. gave a communication complexity lower bound of $\Omega(mk)$ for obtaining any c -approximation for distributed clustering [18]. When the dimension d is constant, this matches the upper bounds on the communication complexity of many distributed algorithms in the literature. However, for $d \in \omega(1)$, there is a gap of $\Theta(d)$. We close this gap using a direct-sum reduction from a recent result on distributed mean estimation [22], proving an $\tilde{\Omega}(mkd)$ lower bound for finding the optimal k -means centers. We show this lower bound holds even when the data satisfies approximation stability or spectral stability, and even when the optimal partition is known up front.

Distributed Clustering under Stability. In the second part of the paper, we improve over the results from Section 3, and in some cases we bypass the hardness result from Section 4, by considering clustering under natural stability notions. Specifically, in Section 5, we provide an algorithm for distributed clustering under $(1 + \alpha, \epsilon)$ -approximation stability using communication $O(mk(d + \log n))$. The high-level structure is similar to the algorithm from Section 3. For k -median and k -means, we show the clustering outputted has $O(\epsilon(1 + \frac{1}{\alpha}))$ error. If we further assume that all optimal clusters are size $\Omega(\epsilon n(1 + \frac{1}{\alpha}))$, and increase communication by a factor of m , then the error drops to $O(\epsilon)$ for k -median. For ℓ_p clustering when $p < \log n$, we show the error is $O(\epsilon(1 + (\frac{18}{\alpha})^p))$. This is the first result for ℓ_p clustering under approximation stability when $2 < p < \log n$, even for a single machine, therefore we improve over Balcan et al. For $p = \infty$ (k -center), we provide an algorithm which outputs the optimal clustering under $(2, 0)$ -approximation stability with $O(mkd)$ communication. Due to the lower bound of Balcan et al. [9], this result is optimal with respect to the value of α . In Section 6, we give an algorithm for distributed clustering under spectral stability. We introduce a distributed algorithm which outputs k centers very close to the optimal centers, using communication $\tilde{O}(mk(d + \log n))$.

Locally Consistent Clustering. Finally, in Section 7, we show that *no* communication is necessary in certain settings. We introduce the notion of a *locally consistent clustering*, motivated by several examples, in which the goal for each machine is to create a k -clustering such that each cluster is a subset of a global optimal cluster. If the data globally satisfies approximation stability, we show that no communication is necessary for each machine to output a locally consistent clustering with small total error.

1.2 Related Work

Centralized Clustering. The first constant-factor approximation algorithm for k -median was given by Charikar et al. [16], and the current best approximation ratio is 2.675 from Byrka et al. [15]. For k -center, there is a tight 2-approximation algorithm [23]. For k -means, the best approximation ratio is 6.357 [2], and Makarychev et al. recently showed a bicriteria algorithm with strong guarantees [30]. For clustering with outliers, there is a 3-approximation algorithm for k -center with z outliers, as well as a bicriteria $4(1 + 1/\epsilon)$ -approximation algorithm for k -median that picks $(1 + \epsilon)z$ outliers [17]. Chen found a true constant factor approximation algorithm for k -median (the constant is not explicitly computed) [19].

Distributed Clustering. Balcan et al. showed a coresets construction for k -median and k -means, which leads to a clustering algorithm with $\tilde{O}(mkd)$ communication, and also studied more general graph topologies for distributed computing [11]. Bateni et al. introduced a construction for *mapping coresets*, which admits a distributed clustering algorithm that can handle balance constraints with communication cost $\tilde{O}(mk)$ [12]. Malkomes et al. showed a distributed 13- and 4- approximation algorithm for k -center with and without outliers, respectively [31]. Chen et al. studied clustering under the broadcast model of distributed computing, and also proved a communication complexity lower bound of $\Omega(mk)$ for distributed clustering [18], building on a recent lower bound for set-disjointness in the message-passing model [14]. Garg et al. showed a communication complexity lower bound for computing the mean of d -dimensional points [22].

Clustering under stability. The notion of approximation stability was defined by Balcan et al. who showed an algorithm which utilizes the structure to output a nearly optimal clustering [8]. Balcan et al. showed an algorithm to exactly cluster k -center instances satisfying $(2, \epsilon)$ -approximation

stability, and they proved a matching lower bound, namely that $(2 - \delta, 0)$ -approximation stability is NP-hard for any δ unless $NP = RP$ [9]. Kumar and Kannan introduced a spectral stability condition which generalized many generative models [27], including Gaussian mixture-models [20, 26], the Planted Partition model [32], as well as deterministic conditions [33]. This work was later improved along several axes, including the dependence on k in the condition, by Awasthi and Sheffet [7]. Chen et al. study distributed algorithms for graph partitioning when the graphs satisfy a notion of stability relating internal expansion of the k pieces to the external expansion [18]. It is known that such graphs also satisfy spectral stability under a suitable Euclidean embedding [6]. See Section 6 for a more detailed comparison of the two notions of stability.

2 Preliminaries

Clustering. Given a set of points V of size n and a distance metric d , let \mathcal{C} denote a clustering of V , which we define as a partition of V into k subsets X_1, \dots, X_k . Each cluster X_i contains a center x_i . When d is an arbitrary distance metric, we must choose the centers from the point set. If $V \subseteq \mathbb{R}^d$ and the distance metric is the standard Euclidean distance, then the centers can be any k points in \mathbb{R}^d . In fact, this distinction only changes the cost of the optimal clustering by at most a factor of 2 when $p = 1, 2$, or ∞ [4]. The ℓ_p cost of \mathcal{C} is

$$\text{cost}(\mathcal{C}) = \left(\sum_i \sum_{v \in X_i} d(v, x_i)^p \right)^{\frac{1}{p}}.$$

We will denote the optimal clustering of a point set V in ℓ_p with z outliers as $\mathcal{OPT}_{k,z,p}(V)$. V , p , k , and/or z will often be clear from context, so we may drop some or all of these parameters. $\mathcal{OPT}(A, B)$ will denote the optimal clustering for a point set $A \subseteq V$, using centers from a different point set $B \subseteq V$. We often overload notation and let \mathcal{OPT} denote the objective value of the optimal clustering as well. In our proofs, we make use of the triangle inequality generalized for ℓ_p , $d(u, v)^p \leq 2^{p-1}(d(u, w)^p + d(w, v)^p)$, for any points u, v, w . We denote the optimal clusters as C_1, \dots, C_k , with centers c_1, \dots, c_k . We say a bicriteria clustering algorithm \mathcal{A} is a (γ, α) -approximation algorithm if it returns $\gamma \cdot k$ centers which define a clustering whose cost is at most an α -factor from the optimal clustering with k centers. Throughout the paper, unless otherwise noted, we assume any d -dimensional datapoint can be expressed using $O(d)$ bits.

Distributed computing. We use a common, general theoretical framework for distributed computing called the *coordinator model*. There are m machines, and machine 1 is designated as the coordinator. Each machine can send messages back and forth with machine 1. This model is very similar to the *message-passing model*, also known as the *point-to-point* model, in which any pair of machines can send messages back and forth. In fact, the two models are equivalent up to small factors in the communication complexity [14]. We assume the data is arbitrarily partitioned across the m machines, and it is the coordinator's job to output the answer. Most of our algorithms can be applied to the mapreduce framework with a constant number of rounds. For more details, see [12, 31].

Communication Complexity. One of our main goals when designing distributed algorithms is to minimize the communication. For an input X and a protocol Π , the *communication cost* is the total number of bits in the messages sent to and from the coordinator. When designing algorithms, we wish to minimize the *communication complexity*, or the maximum communication

cost over all possible inputs X . When proving a lower bound for a problem \mathcal{A} , we define the δ -error communication complexity as the minimum communication complexity of any randomized protocol Π , such that for all inputs X , the probability that Π outputs an incorrect answer is at most δ . For brevity, we use communication complexity to mean .99-error communication complexity.

Approximation Stability. In Sections 4 and 5, we consider clustering under a natural property of real-world instances called *approximation stability*. Intuitively, a clustering instance satisfies this assumption if all clusterings close in *value* to \mathcal{OPT} are also close in terms of the clusters themselves. This is a desirable property when running an approximation algorithm, since in many applications, the k -means or k -median costs are proxies for the final goal of recovering a clustering that is close to the desired “target” clustering. Approximation stability makes this assumption explicit. First we define two clusterings \mathcal{C} and \mathcal{C}' as ϵ -close, if only an ϵ -fraction of the input points are clustered differently in the two clusterings, formally, there exists a permutation σ of $[k]$ such that $\sum_{i=1}^k |C_i \setminus C'_{\sigma(i)}| \leq \epsilon n$.

Definition 1. A clustering instance (V, d) satisfies $(1 + \alpha, \epsilon)$ -approximation stability if all clusterings \mathcal{C} with $\text{cost}(\mathcal{C}) \leq (1 + \alpha) \cdot \mathcal{OPT}$ are ϵ -close to \mathcal{C} .

Spectral Stability. In Sections 4 and 6, we consider clustering under spectral stability. This is a deterministic condition on a clustering dataset that generalizes many assumptions that exist in the study of generative models such as Gaussian Mixture Models. Let A be an $n \times d$ data matrix consisting of n points in a d -dimensional Euclidean space. Let C be an $n \times d$ rank k matrix with each row consisting of the center of the corresponding datapoint.

Definition 2. We say that the matrix A satisfies γ -spectral stability for some constant $\gamma > 0$, if for every $A_i \in C_r$ and for every $r \neq s$, the projection of A_i onto the line joining c_r and c_s is closer to c_r than to c_s by an additive factor of $\left(\frac{\gamma k}{|C_r|} + \frac{\gamma k}{|C_s|}\right) \|A - C\|$.

3 General Robust Distributed Clustering

In this section, we give a general algorithm for distributed clustering with the ℓ_p objective, with or without balance constraints and with or without outliers. This generalizes previous distributed clustering results [12, 31], and answers an open question of Malkomes et al. [31]. We give a simple algorithmic framework, together with a careful analysis, to prove strong guarantees in various settings. Each machine performs a k -clustering on its own data, and the centers, along with the size of their corresponding clusters, are sent to a central machine, which then runs a weighted clustering algorithm on the mk centers (see Figure 1). For the case of clustering with outliers, each machine runs a $(k + z)$ -clustering, and the central machine runs a clustering algorithm that handles outliers.

Theorem 3. Given a sequential (δ, α) -approximation algorithm \mathcal{A} ¹ for balanced k -clustering with the ℓ_p objective with z outliers, and given a sequential (γ, β) -approximation algorithm \mathcal{B} for k -clustering with the ℓ_p objective, then Algorithm 1 is a distributed algorithm for clustering in ℓ_p with z outliers, with communication cost $O(m(k + z)(d + \log n)\gamma)$. The number of centers opened is δk and the approximation ratio is $(2^{3p-1}\alpha^p\beta^p + 2^{2p-1}(\alpha^p + \beta^p))^{1/p}$. For k -median and k -center, this ratio simplifies to $4\alpha\beta + 2\alpha + 2\beta$.

¹ We note that \mathcal{A} must be able to handle weighted points. It has been pointed out that every clustering algorithm we are aware of has this property [12].

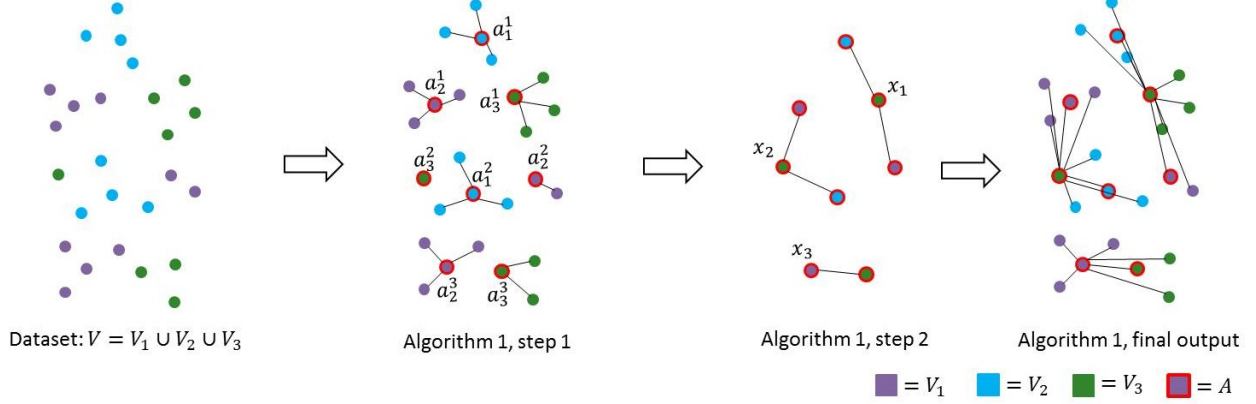


Figure 1: Algorithm 1 for k -median, $m = 3$, $k = 3$.

Algorithm 1 Distributed balanced clustering with outliers

Input: Distributed points $V = V_1 \cup \dots \cup V_m$, algorithms \mathcal{A} and \mathcal{B}

1: For each machine i ,

- Run \mathcal{B} for $(k + z)$ -clustering on V_i , outputting $A_i = \{a_1^i, \dots, a_{k+z}^i\}$.
- Set $w_j^i = |\{p \in V_j \mid a_j^i = \arg\min_{a \in A_i} d(p, a)\}|$.
- Send A_i and all weights to machine 1.

2: Run \mathcal{A} on $\bigcup_i A_i$ using the corresponding weights, outputting $X = \{x_1, \dots, x_k\}$.

Output: Centers $X = x_1, \dots, x_k$

Setting \mathcal{B} to be the $(\frac{8 \log n}{\epsilon}, 1 + \epsilon)$ -bicriteria approximation algorithm for k -median [29], the approximation ratio becomes $6\alpha + 2 + \epsilon$, which improves over the 32α approximation ratio of Bateni et al. [12]. If we set \mathcal{A} as the current best k -median algorithm [15], we achieve a distributed $(18.05 + \epsilon)$ -approximation algorithm for k -median. If instead we plug in the sequential approximation algorithm for k -median with z outliers [19], we obtain the first constant-factor approximation algorithm for k -median with outliers, answering an open question from Malkomes et al. [31]. We can also use the results from Gupta and Tangwongsan [24] to obtain an $O(1)$ -approximation algorithm for $1 < p < \log n$.

Our proof of Theorem 3 carefully reasons about the optimal clustering in certain settings where subsets of the outliers are removed, to ensure the constant approximation guarantee carries through to the final result. First we bound the sum of the local optimal $(k + z)$ -clustering on each machine by the global clustering with outliers in the following lemma (a non-outlier version of this lemma appears in [12]).

Lemma 4. For a partition V_1, \dots, V_k of V and $1 \leq p < \infty$, $\sum_{i=1}^m \text{OPT}_{k+z}(V_i)^p \leq 2^p \text{OPT}_{k,z}^p$.

Proof. Given a machine with datapoints $V_i \subseteq V$, we will first show that

$$\text{OPT}(V_i, V_i)^p \leq 2^p \text{OPT}(V_i, V)^p.$$

Let c_1, \dots, c_k be the optimal centers for $\text{OPT}(V_i, V)$. Given c_j , let c'_j be the closest point in V_i to c_j . Note that there may be one point c' which is the closest point in V_i to two different centers, but this just means we will end up with $\leq k$ centers total, which is okay. Then we have the following:

$$\text{OPT}(V_i, V_i)^p \leq \sum_{v \in V_i} d(v, c'_v)^p$$

$$\begin{aligned}
&\leq 2^{p-1} \sum_{v \in V_i} (d(v, c_v)^p + d(c_v, c'_v)^p) \\
&\leq 2^p \sum_{v \in V_i} d(v, c_v)^p \\
&\leq 2^p \mathcal{OPT}(V_i, V)^p
\end{aligned}$$

The third inequality follows because c'_v was defined as the closest point in V_i to c_v . By choosing $k' = k + z$, we have that $\mathcal{OPT}_{k+z}(V_i, V_i)^p \leq 2^p \mathcal{OPT}_{k+z}(V_i, V)^p$ for all i .

Let the centers in $\mathcal{OPT}_{k,z}$ be c_1, \dots, c_k , and for $v \in V$, let c_v denote the closest of these centers to v . Given the outliers Z from $\mathcal{OPT}_{k,z}$, let $V'_i = V_i \setminus Z$. Then $\mathcal{OPT}_{k+z}(V_i, V)^p \leq \mathcal{OPT}_k(V'_i, V)^p \leq \sum_{v \in V'_i} d(v, c_v)^p$. The second inequality follows because with $k + z$ centers, we can make all points in $V_i \cap Z$ a center and also use the centers in $\mathcal{OPT}_k(V'_i, V)$.

Summing over all i , we arrive at $\sum_i \mathcal{OPT}_{k+z}(V_i, V)^p \leq \mathcal{OPT}_{k,z}^p$, and the lemma follows. \square

Now we prove Theorem 3.

Proof. (Theorem 3) Given a (δ, α) -approximation algorithm \mathcal{A} for balanced clustering in ℓ_p with z outliers and a (γ, β) -approximation algorithm \mathcal{B} for ℓ_p clustering, we show that Algorithm 1 outputs a set X of centers with provable approximation guarantees. First we consider the case where $p < \infty$.

We start by defining all the notation we need for the proof. Let Z denote the set of outliers returned by Algorithm 1 when running \mathcal{B} , let Z^* denote the outliers in $\mathcal{OPT}_{k,z}(A, A)$, where $A = \bigcup_t A_t$ (defined in Algorithm 1), and let Z' denote the outliers in $\mathcal{OPT}_{k,z}$. Denote the centers in $\mathcal{OPT}_{k,z}(A, A)$ by x^* , denote the centers in $\mathcal{OPT}_{k,z}$ by c_j , and let c'_j denote the closest point in A to c_j . Given a point $v \in V_t$, let a_v , x_v , and c_v denote the closest point to v in A_t , X , or $\mathcal{OPT}(V)$, respectively. Finally, let c'_v denote the closest point to c_v in A .

Using the triangle inequality and the fact that for all v , $d(v, x_v) \leq d(v, x_{a_v})$,

$$\sum_{v \in V \setminus Z} d(v, x_v)^p \leq \sum_{v \in V \setminus Z} d(v, x_{a_v})^p \leq 2^{p-1} \sum_{v \in V \setminus Z} (d(v, a_v)^p + d(a_v, x_{a_v})^p) \quad (1)$$

$$\leq 2^{p-1} \sum_{v \in V} d(v, a_v)^p + 2^{p-1} \sum_{v \in V \setminus Z} d(a_v, x_{a_v})^p \quad (2)$$

We can bound the first summation in Expression 2 as follows.

$$\sum_{v \in V} d(v, a_v)^p = \sum_i \sum_{v \in V_i} d(v, a_v)^p \quad (3)$$

$$\leq \sum_i \beta^p \mathcal{OPT}_{k+z}(V_i, V_i)^p \quad [\text{by definition of } \mathcal{B}] \quad (4)$$

$$\leq 2^p \beta^p \mathcal{OPT}_{k,z}^p \quad [\text{by Lemma 4.}] \quad (5)$$

Now we show how to bound the second summation.

$$\begin{aligned}
\sum_{v \in V \setminus Z} d(a_v, x_{a_v})^p &\leq \alpha^p \sum_{v \in V \setminus Z^*} d(a_v, x_v^*)^p && [\text{by def'n of } \mathcal{A}] \\
&\leq \alpha^p \sum_{v \in V \setminus Z'} d(a_v, c'_v)^p && [\text{by def'n of } \mathcal{OPT}_{k,z}(A, A)]
\end{aligned}$$

$$\begin{aligned}
&\leq 2^{p-1}\alpha^p \sum_{v \in V \setminus Z'} (d(a_v, c_v)^p + d(c_v, c'_v)^p) && \text{[by triangle ineq.]} \\
&\leq 2^p \alpha^p \sum_{v \in V \setminus Z'} d(a_v, c_v)^p && \text{[by definition of } c'_v] \\
&\leq 2^{2p-1}\alpha^p \sum_{v \in V \setminus Z'} (d(v, a_v)^p + d(v, c_v)^p) && \text{[by triangle ineq.]} \\
&\leq 2^{2p-1}\alpha^p \sum_{v \in V \setminus Z'} d(v, a_v)^p + 2^{2p-1}\alpha^p \sum_{v \in V \setminus Z'} d(v, c_v)^p && \text{[expanding]} \\
&\leq 2^{3p-1}\alpha^p \beta^p \mathcal{OPT}_{k,z}^p + 2^{2p-1}\alpha^p \mathcal{OPT}_{k,z}^p && \text{[by Expression 5]} \\
&\leq (2^{3p-1}\alpha^p \beta^p + 2^{2p-1}\alpha^p) \mathcal{OPT}_{k,z}^p && \text{[arithmetic]}
\end{aligned}$$

Therefore, our final result is

$$\left(\sum_{v \in V \setminus Z} d(v, x_v)^p \right)^{\frac{1}{p}} \leq (2^{3p-1}\alpha^p \beta^p + 2^{2p-1}(\alpha^p + \beta^p))^{\frac{1}{p}} \mathcal{OPT}_{k,z}.$$

For the communication complexity, it is clear that we communicate $\leq (k+z)m\gamma$ total points and weights, and the weights are numbers $\leq n$, so the communication cost is $O(m(k+z)(d + \log n)\gamma)$. The balance constraints follow from the guarantees of Algorithm \mathcal{A} , and the fact that the points in A are weighted. We also note that for all i , machine 1 can send the center assignments of A_i to machine i , so that each datapoint knows its global center (and this does not increase the communication cost above $O(m(k+z)(d + \log n)\gamma)$).

For k -center, we can derive the same result as k -median by using the same analysis, but replacing every summation with a maximum. For instance, we use a modified Lemma 4 to show $\max_{i=1}^m \mathcal{OPT}_{k+z}(V_i) \leq 2\mathcal{OPT}_{k,z}$. Our final result for k -center is $\max_{v \in V \setminus Z} d(v, x_v) \leq (4\alpha\beta + 2\alpha + 2\beta)\mathcal{OPT}_{k,z}$. \square

4 Communication Complexity Lower Bounds

In this section, we give a dimension-dependent lower bound on the communication complexity for distributed k -means clustering. This generalizes the result of Chen et al. [18], who showed a lower bound independent of the dimension. Our lower bound holds even when the data satisfies $(1 + \alpha, \epsilon)$ -approximation stability or γ -spectral stability for arbitrary α , ϵ , or γ . We also note a corollary from Chen et al. [18].

Corollary 5. *For any $c \geq 1$, $p \in \mathbb{N}$, and $z \geq 0$, the communication complexity of computing a c -approximation for k -clustering in ℓ_p with z outliers is $\Omega(m(k+z))$.*

This follows from Theorem 4.1 in [18] because k -clustering with z outliers is a $k+z$ eligible function: it evaluates to 0 if there are at most $k+z$ points, otherwise it is greater than 0. This shows that for constant dimension, the communication complexity of our Theorem 3 for clustering with outliers is tight up to logarithmic factors.

Now we move on to a dimension-dependent lower bound. Specifically, we lower bound the communication complexity to compute the optimal centers (or the optimal cost) for distributed k -means clustering. Interestingly, this lower bound holds even if the coordinator knows the optimal clusters up front, and just needs to calculate the k different means. Indeed, our method will use a direct-sum theorem on computing the mean of m different data points.

The communication complexity needed to compute the sum of m numbers, each on a different machine, is $\Omega(m)$ [36]. Clearly, the same result holds for averaging m numbers. Now we use a direct-sum theorem to generalize this result to d -dimensional numbers in Euclidean space [22]. The full details are in Appendix A.

Theorem 6. *The communication complexity to compute the optimal clustering for mk points in d dimensions, where each machine contains k points, is $\Omega\left(\frac{mkd}{\log(md)}\right)$ even if the clustering is promised to satisfy $(1 + \alpha, \epsilon)$ -approximation stability for any α, ϵ , or γ -spectral stability for any γ .*

Proof sketch We reduce the problem of computing the optimal clustering across m machines, to the mean estimation problem with m machines, each with one kd -dimensional datapoint. Given such a mean estimation instance, we break up each data point $X \in [-1, 1]^{kd}$ into k “chunks” of d -dimensional data points: for $1 \leq i \leq k$, let $X^i = X_{[1+k(i-1), ki]} \in [-1, 1]^d$. Then we add offsets $4i \cdot [1]^d$ to each vector X^i .

Now we have a set of mk d -dimensional data points, k per machine, defining a clustering instance. We show that any protocol which solves this clustering instance can solve the original mean estimation instance. First we show that the optimal clusters are exactly the k sets of data points Y^i corresponding to all i th chunks of the original data points, which follows because of the added offsets. Therefore, if the coordinator correctly computes the mean of each cluster, it knows the mean of each length- d chunk of dimensions from the original data points. The coordinator can subtract the offsets from the centers and concatenate them, to find the mean of the original data points. This follows because the ℓ_2 mean function is linear across dimensions. We obtain a communication complexity of $\Omega\left(\frac{mkd}{\log(md)}\right)$ by plugging the results of Garg et al. [22] (included in Appendix A) and Viola [36]. To show this result holds under stability, we increase the offsets by factors of $(2(1 + \alpha))$ or $2\gamma \log k$, from which it follows that the clustering instance is stable. \square

5 Distributed Clustering under Approximation Stability

In this section, we improve over the results from Section 3 if it is known up front that the data satisfies approximation stability. We give modified algorithms which leverage this structure to output a clustering very close to optimal, using no more communication than the algorithm from Section 3. We focus on k -median, but we give the details for k -center and ℓ_p for $p < \log n$ in the appendix.

We give a two-phase algorithm for k -median under approximation stability. The high level structure of the algorithm is similar to Algorithm 1: first each machine clusters its local point set, and sends the weighted centers to the coordinator. Then the coordinator runs a weighted clustering algorithm to output the solution. The difference lies in the algorithm that is run by the machines and the coordinator, which will now take advantage of the approximation stability assumption. We obtain the following result.

Theorem 7. *Algorithm 3 outputs a set of centers defining a clustering that is $O(\epsilon(1 + \frac{1}{\alpha}))$ -close to OPT for k -median under $(1 + \alpha, \epsilon)$ -approximation stability with $O(mk(d + \log n))$ communication.*

We achieve a similar result for k -means. For clustering in ℓ_p when $p < \log n$, the error of the outputted clustering increases to $O(\epsilon(1 + (\frac{18}{\alpha})^p))$ (Theorem 15). This is the first result for ℓ_p clustering under approximation stability when $2 < p < \log n$, even for a single machine, thereby improving over Balcan et al. We also show that if the optimal clusters are not too small, the error of the outputted clustering can be pushed even lower.

Algorithm 2 Iterative greedy procedure

Input: Graph $G = (V, E)$, parameter k

- 1: Initialize $A = \emptyset$, $V' = V$.
- 2: While $|A| < k$, set $v' = \max_{v \in V'} N(v) \cap V'$, $C(v') = N(v') \cap V'$.
 - Add $(v', C(v'))$ to A , and remove $N(v')$ from V' .

Output: Center and cluster pairs $A = \{(v_1, C(v_1)), \dots, (v_k, C(v_k))\}$

Algorithm 3 Distributed k -median clustering under $(1 + \alpha, \epsilon)$ -approximation stability

Input: Distributed points $V = V_1 \cup \dots \cup V_m$, average k -median cost w_{avg}

- 1: Set $t = \frac{\alpha w_{avg}}{18\epsilon}$
- 2: For each machine i ,
 - Create the threshold graph G_{2t}^i using distance $2t$.
 - Run Algorithm 2 with input (G_{2t}^i, k) , outputting $A'_i = \{(v_1^i, C(v_1^i)), \dots, (v_k^i, C(v_k^i))\}$.
 - Send $A_i = \{(v_1^i, |C(v_1^i)|), \dots, (v_k^i, |C(v_k^i)|)\}$ to Machine 1.
- 3: Given the set of weighted points received, $A = \cup_i A_i$, create the threshold graph G_{6t} using points A and distance $6t$.
- 4: Run Algorithm 2 with graph G_{6t} (using weighted points) and parameter k , outputting $X' = \{(x_1, C(x_1)), \dots, (x_k, C(x_k))\}$.

Output: Centers $X = \{x_1, \dots, x_k\}$

Theorem 8. *There exists an algorithm which outputs a clustering that is $O(\epsilon)$ -close to \mathcal{OPT} for k -median under $(1 + \alpha, \epsilon)$ -approximation stability with $O(m^2kd + mk \log n)$ communication if each optimal cluster C_i has size $\Omega(\epsilon n(1 + \frac{1}{\alpha}))$.*

We start by stating two properties which demonstrate the power of approximation stability. Let w_{avg} denote the average distance from each point to its optimal center, so $w_{avg} \cdot n = \mathcal{OPT}$.

Lemma 9. [8] *Given a $(1 + \alpha, \epsilon)$ -approximation stable clustering instance (V, d) for k -median, then*

Property 1: *For all x , for $\leq \frac{\alpha \epsilon n}{\alpha}$ points v , $d(v, c_v) \geq \frac{\alpha w_{avg}}{36\epsilon}$.*

Property 2: *For $< 6\epsilon n$ points v , $\exists c_i \neq c_v$ such that $d(v, c_i) - d(v, c_v) \leq \frac{\alpha w_{avg}}{2\epsilon}$.*

Property 1 follows from Markov's inequality. Property 2 follows from the $(1 + \alpha, \epsilon)$ -approximation stability condition: If more than $6\epsilon n$ points are almost the same distance to their second-closest center as their closest center, then we can assign these points to their second-closest center, achieving a low-cost clustering that is not ϵ -close to \mathcal{OPT} , contradicting approximation stability.

Now we define a point as *bad* if it falls into the bad case of either Property 1 or Property 2 with $x = 36$. Formally, $B = \{v \mid d(v, c_v) \geq \frac{\alpha w_{avg}}{36\epsilon} \text{ or } \exists c_i \neq c_v \text{ s.t. } d(v, c_i) - d(v, c_v) \leq \frac{\alpha w_{avg}}{2\epsilon}\}$. From Lemma 9, $|B| \leq (6 + \frac{36}{\alpha})\epsilon n$. Otherwise, a point is *good*. Given an optimal cluster C_i , define $H_i = C_i \setminus B$, the set of good points in C_i . Given a set of points V with a distance metric d and $t > 0$, we define the threshold graph G_t as a graph on V , where there is an edge between each pair $u, v \in V$ if and only if $d(u, v) \leq t$. We use this concept in our algorithm.

We prove the following about Algorithm 2 (formally, Lemma 14 in Appendix B): Given the input graph G contains a partition A_1, \dots, A_k, A' with the following guarantee: *Condition (1)* for all i , for all $u, v \in A_i$, $(u, v) \in E(G)$, and *Condition (2)* for all $i \neq j$, $u \in A_i$, and $v \in A_j$, then $(u, v) \notin E(G)$, moreover, u and v do not share a common neighbor in G . Then Algorithm 2 outputs

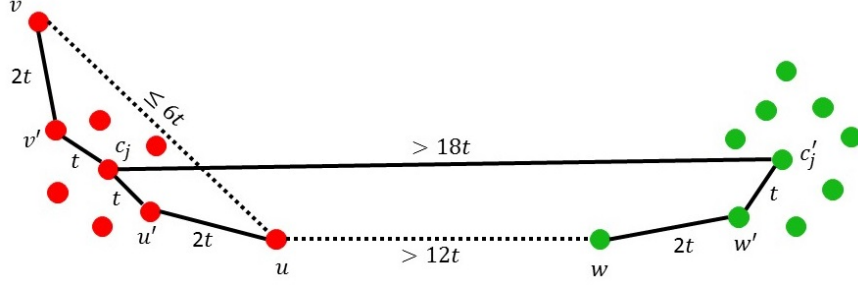


Figure 2: Proof of Theorem 7

a clustering that is $3|A'|$ -close to A_1, \dots, A_k . Now we can prove Theorems 7 and 8. The full details are in Appendix B.

Proof sketch (Theorem 7) We assume Algorithm 3 knows the value w_{avg} , but in the appendix, we show how to relax this condition. First, given machine i , let $\{H_1^i, \dots, H_k^i\}$ denote the good clusters, and let B_i denote the set of bad points on machine i . We use Lemma 9 to show that in graph G_{2t}^i , the good clusters satisfy Conditions (1) and (2). Therefore, by Lemma 14, the clustering outputted in Step 2 is $3|B_i|$ -close to the good clusters $\{H_1^i, \dots, H_k^i\}$. The total error over all machines is $< 3|B|$, and it follows that all but $< 3|B|$ good points are within $2t$ of some point in A .

Now, we partition A into sets H_1^A, \dots, H_k^A, B' , where H_j^A denotes points which are distance $2t$ to good points from H_j , and B' contains points far from all good points. This partition is well-defined because any pair of good points from different clusters are far apart. From the previous paragraph, $|B'| \leq 3|B|$ (we let $|B'|$ denote the sum of the weights of all points in B'). Again using Lemma 9, we show that H_1^A, \dots, H_k^A, B' in graph G_{6t} satisfy Conditions (1) and (2). Given two points $u, v \in H_j^A$, then there exist $u', v' \in H_j$ such that $d(u, v) \leq d(u, u') + d(u', c_j) + d(c_j, v') + d(v', v) \leq 6t$. Given $u \in H_j^A$ and $w \in H_{j'}^A$, there exist $u' \in H_j, w' \in H_{j'}$ such that $d(u', c_{j'}) > 18t - d(c_j, u')$, which we use to show u and w cannot have a common neighbor in G_{6t} (see Figure 2).

Therefore, by Lemma 14, the clustering outputted in Step 4 is $3|B'|$ -close to the good clusters $\{H_1^A, \dots, H_k^A\}$. It follows that there exists a bijection between each center outputted x_j , and each good cluster H_j^A , and all but $3|B'| \leq 9|B|$ good points $v \in H_j$ are distance $2t$ to a point in A which is distance $6t$ to x_j , and v must be distance $> 8t$ to any other outputted center from Lemma 9. So the error over all points, good and bad, is $9|B| + |B| = 10|B| \in O(\epsilon n(1 + \frac{1}{\alpha}))$, so the algorithm achieves the desired error bound. The total communication is mk points and mk weights, or $O(mk(d + \log n))$ bits. \square

Proof sketch (Theorem 8) The algorithm is as follows. First, run Algorithm 3. Then send X' to each machine i , incurring a communication cost of $O(m^2kd)$. For each point $v \in V$, calculate the median distance from v to each cluster $C(x_j)$ (using the weights), and assign v to the index j with the minimum median distance. Call the new clusters X_1, \dots, X_k , according to the indices. We will prove this clustering is $O(\epsilon)$ -close to the optimal clustering. Specifically, we will show that all points are correct except for the $6\epsilon n$ points in the bad case of Property 2 from Lemma 9.

From the proof of Theorem 7, we know that at most $3|B|$ points from the clustering $\{H_1^A, \dots, H_k^A\}$ are misclassified with respect to H_1, \dots, H_k , so all but $3|B| \in O(\epsilon n(1 + \frac{1}{\alpha}))$ weighted points $v \in H_j^A$ are within $2t$ of a point in H_j . By assumption, all clusters H_j^A consist of more than half of these good (proxy) points. Given a point $v \in C_j$ satisfying the good case in Property 2, we use the triangle inequality to show $d(v, u) \leq d(v, c_j) + 3t$ for all $u \in H_j^A$, and $d(v, u) > d(v, c_j) + 15t$ for all $u \in H_{j'}^A$, therefore, v will be assigned to the correct cluster. \square

For k -center, we show an algorithm that outputs the exact solution under $(2, 0)$ -approximation stability using $O(mkd)$ communication. Even in the single machine setting, it is NP-hard to find the exact solution to k -center under $(2 - \epsilon, 0)$ -approximation stability unless $NP = RP$ [9], therefore our result is tight with respect to the level of stability. We include all details in Appendix B.

6 Distributed Clustering under Spectral Stability

In this section, we give a distributed clustering algorithm under spectral stability. Recall the definition of matrices A and C from Section 2. Our distributed algorithm for spectral stability works as follows. We first compute a k -SVD of the data matrix and project the data onto the span of the top k singular vectors. This can be done in a distributed manner [28]. We then run a distributed constant factor approximation algorithm for k -means developed in Section 3. Finally, we run a natural distributed version of the popular Lloyd's heuristic for a few rounds to converge to nearly optimal centers. We achieve the following theorem.

Theorem 10. *Let A be a data matrix satisfying γ -spectral stability. Then, Algorithm 4 on A outputs centers $\nu_1, \nu_2, \dots, \nu_k$ on machine 1 such that $\|\nu_i - c_i\| \leq \epsilon$. The total communication cost is $O(mk(d + \log n) \log(\frac{k}{\epsilon}))$.*

Proof. At each step of the for loop, every machine is computing a local weighted mean for each of its k clusters. Given such information from each machine, machine 1 can easily update the means to the new values exactly as in the Lloyd's step. The correctness of the algorithm then follows from previous work [7, 27] that shows that in $T = \log(\frac{\|A\|}{\epsilon})$ steps the cluster centers will be recovered to ϵ accuracy. To bound the communication cost, the result of Balcan et al. [28] bounds the communication cost of computing k -SVD to be $O(mkd)$. From Theorem 3, Algorithm 1 uses $O(mk(d + \log n))$ bits of communication. Finally, in each iteration of the distributed Lloyd's algorithm, each machine receives k data points and sends out k data points along with k weight values. Hence, the communication cost per iteration is $O(mk(d + \log n))$. Combining, and because $\log \|A\| \in O(\log k)$, we get that the overall communication cost is $O(mk(d + \log n) \log(\frac{k}{\epsilon}))$. \square

Algorithm 4 Distributed Spectral Clustering

Input: $n \times d$ data matrix A distributed over m machines, parameter k , accuracy ϵ .

- 1: Run the distributed algorithm from [28] to compute \hat{A}_i 's, i.e., the projection of A_i onto the top k singular vectors of A .
- 2: Run Algorithm 1 for k -means to compute initial centers ν_1^0, \dots, ν_k^0 . Set $T = \log(\frac{\|A\|}{\epsilon})$.
- 3: For $t = 1$ to T ,
 - Machine 1 sends $\nu_1^{t-1}, \dots, \nu_k^{t-1}$ to every other machine.
 - For each machine i ,
 - Compute local clustering using $\nu_1^{t-1}, \dots, \nu_k^{t-1}$
 - Compute the mean $\mu_{i,j}^{t-1}$ and the weight $w_{i,j}^{t-1}$ for each cluster
 - Send all $\mu_{i,j}^{t-1}$ and $w_{i,j}^{t-1}$ to machine 1.
 - Machine 1 updates $\nu_j^t = \frac{\sum_{i=1}^m w_{i,j}^{t-1} \mu_{i,j}^{t-1}}{\sum_{i=1}^m w_{i,j}^{t-1}}$ for $j = 1, 2, \dots, k$.
- 4: Output ν_1^T, \dots, ν_k^T .

Output: Centers $\nu_1, \nu_2, \dots, \nu_k$

A notion related to spectral stability has been recently studied [18] for distributed graph partitioning. Given a graph on n vertices with the ground truth partitioning as V_1, V_2, \dots, V_k , let $\rho(k)$ be the maximum edge expansion of any piece, and let $\lambda_{k+1}(G)$ be the $k+1$ th smallest eigenvalue of the normalized graph Laplacian. Then the graph is stable if $\frac{\lambda_{k+1}(L_G)}{\rho_k} = \Omega(k^3)$. Chen et al. design communication efficient distributed algorithms to cluster such stable graphs [18]. These graphs are intimately connected to spectral stability. In fact, it was shown that stable graphs in the above sense also satisfy spectral stability under an appropriate Euclidean embedding of the nodes of the graph [6]. Hence, in principle, our distributed algorithm for spectral stability can also be applicable to cluster stable graphs. However, Chen et al. study a setting where the edges are partitioned across machines and hence their result is formally incomparable to ours [18].

7 Locally Consistent Clustering

In this section, we show how to output a nearly optimal *local* clustering with no communication, provided the data satisfies approximation stability globally. First we provide a few motivating examples. Given m different hospitals, each of which have data on their own patients, each patient is at low or high risk for certain diseases, and there exists an (unknown) ground truth k -clustering of the patients corresponding to their risks for the diseases. We want to conduct a representative survey, with at least one patient from each cluster and each hospital. One solution is to run a distributed clustering algorithm to approximate the ground truth clustering using $\Omega(mkd)$ communication, and then randomly sample m patients from each cluster. However, in Theorem 12 we show that if the data is globally structured, we can run a clustering algorithm on each individual hospital, then sample one patient from each cluster at each hospital.

Another example is when each hospital wants to decide, for each pair of patients, whether one can donate blood to the other. In this application, the global clustering is irrelevant – the only important information is pairwise information. Again, assume there is a ground truth clustering, and one patient can donate to another if and only if they are in the same cluster. If the data satisfies approximation stability, we can cluster locally, requiring no communication.

Definition 11. *Given a clustering instance (V, d) with optimal clustering \mathcal{OPT} , assuming the data is distributed across m machines, an ϵ -locally consistent clustering is a k -clustering C_1^i, \dots, C_k^i for each machine i , such that the global clustering $\mathcal{C} = \{\bigcup_{j=1}^k C_j^i \mid i \in [m]\}$ is ϵ -close to \mathcal{OPT} .*

In words, an ϵ -locally consistent clustering is a clustering for each machine, such that for each machine i and optimal cluster j , machine i contains a cluster that is nearly a subset of C_j – the total error across all pairs $1 \leq i \leq m$ and $1 \leq j \leq k$ is ϵ .

Theorem 12. *Given a k -median instance satisfying $(1 + \alpha, \epsilon)$ -approximation stability, then there is an efficient algorithm which outputs an $O(\epsilon(1 + \frac{1}{\alpha}))$ -locally consistent clustering. The communication complexity is 0.*

Proof. Each machine runs steps 1 and 2 from Algorithm 3, i.e. the algorithm from Balcan et al. [8].

We will refer to the details of Theorem 7. Recall in Theorem 7, for each machine i , and $1 \leq j \leq k$, let H_j^i denote the set of good points from cluster C_j on machine i . Let B_i denote the set of bad points on machine i . Then after Steps 1 and 2 from Algorithm 3, machine i outputs k clusters $C(v_j)$ such that there exists a bijection $\sigma : [k] \rightarrow [k]$ between the clusters $C(v_j)$ and the good clusters H_ℓ^i such that $\sum_j |H_{\sigma(j)}^i \setminus C(v_j)| \leq 4|B_i|$. For ease of notation, we now assume that σ is the identity bijection for all machines.

Since $H_j = \cup_i H_j^i$, then for the clusters $X_j = \cup_i C(v_j)$, we have $\sum_j |H_j \setminus X_j| = \sum_j |\cup_i H_j^i \setminus \cup_i C(v_j)| = \sum_j |\cup_i (H_j^i \setminus C(v_j))| = \sum_i \sum_j |H_j^i \setminus C(v_j)| \leq \sum_i |B_i| \leq |B| \in O(\epsilon n(1 + \frac{1}{\alpha}))$. This completes the proof. \square

8 Conclusion

We present a simple and general framework for distributed clustering with outliers. We give an algorithm for k -clustering for any ℓ_p objective with z outliers using $\tilde{O}(m(k+z)(d+\log n))$ bits of communication, answering an open question [31] and improving over the previous best approximation ratio [12]. For distributed k -means clustering, we give the first dimension-dependent communication complexity lower bound for finding the optimal clustering, improving over the lower bound of Chen et al. [18]. Our lower bound holds even when the data is stable. An interesting open question is to extend this result to any ℓ_p objective.

We show how to improve the quality of the clustering produced, provided the data satisfies certain natural notions of stability, specifically, approximation stability and spectral stability [8, 27]. In certain applications that only require locally consistent clusterings, we show that no communication is necessary if the data satisfies approximation stability. It is an interesting question to study the locally consistent clustering problem in different settings.

References

- [1] Manu Agarwal, Ragesh Jaiswal, and Arindam Pal. k -means++ under approximation stability. *Theoretical Computer Science*, 588:37–51, 2015.
- [2] Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, and Justin Ward. Better guarantees for k -means and euclidean k -median by primal-dual algorithms. *arXiv preprint arXiv:1612.07925*, 2016.
- [3] Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. Local search heuristics for k -median and facility location problems. *SIAM Journal on Computing*, 33(3):544–562, 2004.
- [4] Pranjali Awasthi and Maria-Florina Balcan. Center based clustering: A foundational perspective. 2014.
- [5] Pranjali Awasthi, Avrim Blum, and Or Sheffet. Center-based clustering under perturbation stability. *Information Processing Letters*, 112(1):49–54, 2012.
- [6] Pranjali Awasthi, Moses Charikar, Ravishankar Krishnaswamy, and Ali Kemal Sinop. Spectral embedding of k -cliques, graph partitioning and k -means. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 301–310. ACM, 2016.
- [7] Pranjali Awasthi and Or Sheffet. Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 37–49. Springer, 2012.
- [8] Maria-Florina Balcan, Avrim Blum, and Anupam Gupta. Clustering under approximation stability. *Journal of the ACM (JACM)*, 60(2):8, 2013.

- [9] Maria-Florina Balcan, Nika Haghtalab, and Colin White. k -center clustering under perturbation resilience. In *Proceedings of the Annual International Colloquium on Automata, Languages, and Programming (ICALP)*, 2016.
- [10] Maria Florina Balcan and Yingyu Liang. Clustering under perturbation resilience. *SIAM Journal on Computing*, 45(1):102–155, 2016.
- [11] Maria-Florina F Balcan, Steven Ehrlich, and Yingyu Liang. Distributed k -means and k -median clustering on general topologies. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1995–2003, 2013.
- [12] Mohammadhossein Bateni, Aditya Bhaskara, Silvio Lattanzi, and Vahab Mirrokni. Distributed balanced clustering via mapping coresets. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2591–2599, 2014.
- [13] Yonatan Bilu and Nathan Linial. Are stable instances easy? *Combinatorics, Probability and Computing*, 21(05):643–660, 2012.
- [14] Mark Braverman, Faith Ellen, Rotem Oshman, Toniann Pitassi, and Vinod Vaikuntanathan. A tight bound for set disjointness in the message-passing model. In *Proceedings of the Annual Symposium on Foundations of Computer Science (FOCS)*, pages 668–677. IEEE, 2013.
- [15] Jarosław Byrka, Thomas Pensyl, Bartosz Rybicki, Aravind Srinivasan, and Khoa Trinh. An improved approximation for k -median, and positive correlation in budgeted optimization. In *Proceedings of the Annual Symposium on Discrete Algorithms (SODA)*, pages 737–756. SIAM, 2015.
- [16] Moses Charikar, Sudipto Guha, Éva Tardos, and David B Shmoys. A constant-factor approximation algorithm for the k -median problem. In *Proceedings of the Annual Symposium on Theory of Computing (STOC)*, pages 1–10. ACM, 1999.
- [17] Moses Charikar, Samir Khuller, David M Mount, and Giri Narasimhan. Algorithms for facility location problems with outliers. In *Proceedings of the Annual Symposium on Discrete Algorithms (SODA)*, pages 642–651. Society for Industrial and Applied Mathematics, 2001.
- [18] Jiecao Chen, He Sun, David Woodruff, and Qin Zhang. Communication-optimal distributed clustering. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pages 3720–3728, 2016.
- [19] Ke Chen. A constant factor approximation algorithm for k -median clustering with outliers. In *Proceedings of the Annual Symposium on Discrete Algorithms (SODA)*, volume 8, pages 826–835, 2008.
- [20] Sanjoy Dasgupta and Leonard Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *Journal of Machine Learning Research*, 8:203–226, 2007.
- [21] Travis Dick, Mu Li, Venkata Krishna Pillutla, Colin White, Maria Florina Balcan, and Alex Smola. Data driven resource allocation for distributed learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.
- [22] Ankit Garg, Tengyu Ma, and Huy Nguyen. On communication cost of distributed statistical estimation and dimensionality. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2726–2734, 2014.

- [23] Teofilo F Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.
- [24] Anupam Gupta and Kanat Tangwongsan. Simpler analyses of local search algorithms for facility location. *arXiv preprint arXiv:0809.2554*, 2008.
- [25] Rishi Gupta, Tim Roughgarden, and C Seshadhri. Decompositions of triangle-dense graphs. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 471–482. ACM, 2014.
- [26] Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. *SIAM Journal on Computing*, 38(3):1141–1156, 2008.
- [27] Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *Proceedings of the Annual Symposium on Foundations of Computer Science (FOCS)*, pages 299–308. IEEE, 2010.
- [28] Yingyu Liang, Maria-Florina F Balcan, Vandana Kanchanapally, and David Woodruff. Improved distributed principal component analysis. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pages 3113–3121, 2014.
- [29] Jyh-Han Lin and Jeffrey Scott Vitter. Approximation algorithms for geometric median problems. *Information Processing Letters*, 44(5):245–249, 1992.
- [30] Konstantin Makarychev, Yury Makarychev, Maxim Sviridenko, and Justin Ward. A bi-criteria approximation algorithm for k means. In *APPROX*, 2016.
- [31] Gustavo Malkomes, Matt J Kusner, Wenlin Chen, Kilian Q Weinberger, and Benjamin Moseley. Fast distributed k-center clustering with outliers on massive data. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1063–1071, 2015.
- [32] Frank McSherry. Spectral partitioning of random graphs. In *Proceedings of the Annual Symposium on Foundations of Computer Science (FOCS)*, pages 529–537. IEEE, 2001.
- [33] Rafail Ostrovsky, Yuval Rabani, Leonard J Schulman, and Chaitanya Swamy. The effectiveness of lloyd-type methods for the k-means problem. *Journal of the ACM (JACM)*, 59(6):28, 2012.
- [34] Ali Kemal Sinop. How to round subspaces: A new spectral clustering algorithm. In *Proceedings of the Annual Symposium on Discrete Algorithms (SODA)*, pages 1832–1847. SIAM, 2016.
- [35] Cheng Tang and Claire Monteleoni. On lloyds algorithm: new theoretical insights for clustering in practice. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS*, pages 9–11, 2016.
- [36] Emanuele Viola. The communication complexity of addition. *Combinatorica*, 35(6):703–747, 2015.
- [37] Konstantin Voevodski, Maria-Florina Balcan, Heiko Röglin, Shang-Hua Teng, and Yu Xia. Min-sum clustering of protein sequences with limited distance information. In *International Workshop on Similarity-Based Pattern Recognition*, pages 192–206. Springer, 2011.

A Proof from Section 4

To prove Theorem 6, we utilize the following direct-sum theorem.

Theorem 13. [22]² Suppose Π computes the mean of m datapoints in $X \in [-1, 1]^d$ distributed across m machines. Then the communication cost is $\Omega\left(\frac{dm}{\log m}\right)$.

Theorem 6 (restated). The communication complexity to compute the optimal clustering for mk points in d dimensions, where each machine contains k points, is $\Omega\left(\frac{mkd}{\log(md)}\right)$ even if the clustering is promised to satisfy $(1 + \alpha, \epsilon)$ -approximation stability for any α, ϵ , or γ -spectral stability for any γ .

Proof. Given a protocol Π for computing the optimal clustering over a distributed clustering instance with B bits of communication, and an instance \mathcal{X} of the mean estimation problem with m machines each with one kd -dimensional datapoint, then we will solve the mean estimation problem using Π .

For each sample point $X \in [-1, 1]^{kd}$, we break it up into k “chunks” $X^1, \dots, X^k \in [-1, 1]^d$, where $X^i = X_{[1+k(i-1), ki]}$. Then we add offsets to each of the new vectors, $Y^i = X^i + 4i \cdot [1]^d$. Denote the set of all datapoints created from chunk i to be \mathcal{Y}^i .

Define the new set $\mathcal{Y} = \cup_i \mathcal{Y}^i$ of km datapoints in d dimensions as a clustering input to protocol Π . By assumption, Π returns the optimal centers c_1, \dots, c_k with B bits of communication. First, we show that the optimal clusters are exactly the k sets of datapoints $\mathcal{Y}^1, \dots, \mathcal{Y}^k$. This is because of the offset we added, which implies each point $p \in \mathcal{Y}^i$ is closer to all other points in \mathcal{Y}^i (distance $\leq 2\sqrt{d}$) than to any point from a set $j \neq i$ (distance $\geq 2\sqrt{d}$). Therefore, the optimal centers c_1, \dots, c_k must be equal to the means of the sets $\mathcal{Y}^1, \dots, \mathcal{Y}^k$.

The coordinator knows the centers at the end of Π . Now, for each center c_i , the coordinator subtracts the offset, $c'_i = c_i - 4i \cdot [1]^d$, to obtain the means for the original data chunks X^i . Finally, the coordinator concatenates the new centers together into a single vector of size kd , $C = [c'_1 \ c'_2 \ \dots \ c'_k]$. This is equal to the mean of the original mean estimation problem, since the mean function is linear over dimensions. Then we obtain a communication complexity of $\Omega\left(\frac{mkd}{\log md}\right)$ by plugging in Theorem 13 and the result of Viola [36].

To obtain the result for clustering under $(1 + \alpha, \epsilon)$ -approximation stability, we increase the offsets by a factor of (2α) , to $8\alpha i \cdot [1]^d$. This ensures that each datapoint is an α -factor closer to its center than to any other point from another cluster, so the data easily satisfies $(\alpha, 0)$ -approximation stability. Similarly, for spectral stability, we increase the offset to $4\|A\| \cdot i$, which easily satisfies the condition. \square

B Proofs from Section 5

Lemma 14. [8]³ Given a graph G over good clusters X_1, \dots, X_k and bad points B , with the following properties:

1. For all u, v in the same X_i , edge (u, v) is in $E(G)$.

² The theorem is in terms of computing the true mean of a Gaussian distribution, but the proof applies for any function $f : [-1, 1]^d \rightarrow \mathbb{R}$ which is linear over the dimensions (such as the mean function).

³This lemma is obtained by merging Lemma 3.6 and Theorem 3.9 [8].

2. For $u \in X_i$, $v \in X_j$ such that $i \neq j$, then $(u, v) \notin E(G)$, moreover, u and v do not share a common neighbor in G .

Then let $C(v_1), \dots, C(v_k)$ denote the output of running Algorithm 2 on G with parameter k . There exists a bijection $\sigma : [k] \rightarrow [k]$ between the clusters $C(v_i)$ and X_j such that $\sum_i |X_{\sigma(i)} \setminus C(v_i)| \leq 3|B|$.

Proof. From the first assumption, each good cluster X_i is a clique in G . Initially, let each clique X_i be “unmarked”, and then we “mark” it the first time the algorithm picks a $C(v_j)$ that intersects X_i . A cluster $C(v_j)$ can intersect at most one X_i because of the second assumption. During the algorithm, there will be two cases to consider. If the cluster $C(v_j)$ intersects an unmarked clique X_i , then set $\sigma(j) = i$. Denote $|X_i \setminus C(v_j)| = r_j$. Since the algorithm chose the maximum degree node and X_i is a clique, then there must be at least r_j points from B in $C(v_j)$. So for all cliques X_i corresponding to the first case, we have $\sum_j |X_{\sigma(j)} \setminus C(v_j)| \leq \sum_j r_j \leq |B|$.

If the cluster $C(v_j)$ intersects a marked clique, then assign $\sigma(j)$ to an arbitrary $X_{i'}$ that is not marked by the end of the algorithm. The total number of points in all such $C(v_j)$ ’s is at most the number of points remaining from the marked cliques, which we previously bounded by $|B|$, plus up to $|B|$ more points from the bad points. Because the algorithm chose the highest degree nodes in each step, each $X_{i'}$ has size at most the size of its corresponding $C(v_j)$. Therefore, for all cliques $X_{i'}$ corresponding to the second case, we have $\sum_j |X_{\sigma(j)} \setminus C(v_j)| \leq \sum_j |X_{\sigma(j)}| \leq 2|B|$. Thus, over both cases, we reach a total error of $3|B|$. \square

Theorem 7 (restated). *Algorithm 3 outputs a set of centers defining a clustering that is $O(\epsilon(1 + \frac{1}{\alpha}))$ -close to \mathcal{OPT} for k -median under $(1 + \alpha, \epsilon)$ -approximation stability with $O(mk(d + \log n))$ communication.*

Proof. The proof is split into two parts, both of which utilize Lemma 14. First, given machine i and $1 \leq j \leq k$, let H_j^i denote the set of good points from cluster C_j on machine i . Let B_i denote the set of bad points on machine i . Given $u, v \in H_j^i$, $d(u, v) \leq d(u, c_j) + d(c_j, v) \leq 2t$, so H_j^i is a clique in G_{2t}^i . Given $u \in H_j^i$ and $v \in H_{j'}^i$, such that $j \neq j'$, then

$$d(u, v) > d(u, c_{j'}) - d(c_{j'}, v) \geq 18t - d(u, c_j) - d(c_j, v) > 16t.$$

Therefore, if u and v had a common neighbor w in G_{2t}^i , $16t < d(u, v) \leq d(u, w) + d(v, w) \leq 4t$, causing a contradiction. Since G_{2t}^i satisfies the conditions of Lemma 14, it follows that there exists a bijection $\sigma : [k] \rightarrow [k]$ between the clusters $C(v_j)$ and the good clusters H_ℓ such that $\sum_j |H_{\sigma(j)}^i \setminus C(v_j)| \leq 3|B_i|$. Therefore, all but $3|B_i|$ good points on machine i are within $2t$ of some point in A_i . Across all machines, $\sum_i |B_i| \leq |B|$, so there are less than $4|B|$ good points which are not distance $2t$ to some point in A .

Since two points $u \in H_i$, $v \in H_j$ for $i \neq j$ are distance $> 16t$, then each point in A is distance $\leq 2t$ from good points in at most one set H_i . Then we can partition A into sets H_1^A, \dots, H_k^A, B' , such that for each point $u \in H_i^A$, there exists a point $v \in H_i$ such that $d(u, v) \leq 2t$. The set B' consists of points which are not $2t$ from any good point. From the previous paragraph, $|B'| \leq 3|B|$, where $|B'|$ denotes the sum of the weights of all points in B' . Now, given $u, v \in H_i^A$, there exist $u', v' \in H_i$ such that $d(u, u') \leq 2t$ and $d(v, v') \leq 2t$, and $d(u, v) \leq d(u, u') + d(u', c_i) + d(c_i, v') + d(v', v) \leq 6t$. Given $u \in H_i^A$ and $w \in H_j^A$ for $i \neq j$, there exist $u' \in H_i$, $w' \in H_j$ such that $d(u, u') \leq 2t$ and $d(w, w') \leq 2t$.

$$d(u, w) \geq d(u', c_j) - d(u, u') - d(c_j, w') - d(w, w') > (18t - d(u, c_i)) - 2t - t - 2t \geq 12t.$$

See Figure 2. Therefore, if u and w had a common neighbor w in G_{6t} , then $12t < d(u, v) \leq d(u, w) + d(v, w) \leq 12t$, causing a contradiction. Since G_{6t} satisfies the conditions of Lemma 14 it follows that there exists a bijection $\sigma : [k] \rightarrow [k]$ between the clusters $C(v_i)$ and the good clusters H_j^A such that $\sum_j |H_{\sigma(j)}^A \setminus C(v_j)| \leq 3|B'|$. Recall the centers chosen by the algorithm are labeled as the set X . Let $x_i \in X$ denote the center for the cluster H_i according to σ . Then all but $3|B'|$ good points $u \in H_i$ are distance $2t$ to a point in A which is distance $6t$ to x_i . u must be distance $> 8t$ to all other points in X because they are distance $2t$ from good points in other clusters. Therefore, all but $3|B'| \leq 12|B|$ good points are correctly clustered. The total error over good and bad points is then $12|B| + |B| = 13|B| \leq (48 + \frac{468}{\alpha})\epsilon n$ so the algorithm achieves error $O(\epsilon(1 + \frac{1}{\alpha}))$. There are mk points communicated to M_1 , and the weights have $\log n$ bits, so the total communication is $O(mk(d + \log n))$. This completes the proof for k -median when the algorithm knows w_{avg} up front.

When Algorithm 3 does not know w_{avg} , then it first runs Algorithm 1 to obtain an estimate $\hat{w} \in [w_{avg}, \beta w_{avg}]$ for $\beta \in O(1)$. As mentioned in Section 3, β can be as low as $18.05 + \epsilon$. Now we reset t in Algorithm 3 to be $\hat{t} = \frac{\alpha\beta w_{avg}}{18\epsilon}$. Then the set of bad points grows by a factor of β , but the same analysis still holds, in particular, Lemma 14 and the above paragraphs go through, adding a factor of β to the error.

As in Theorem 3, machine 1 can send the center assignments of A_i to machine i , so that each datapoint knows its global center, without increasing the bound on the communication cost. \square

Now we show how to generalize Theorem 7 to any ℓ_p objective for $p < \log n$.

Theorem 15. *Given $p < \log n$, there exists an algorithm which outputs a set of centers defining a clustering that is $O(\epsilon(1 + (\frac{18}{\alpha})^p))$ -close to \mathcal{OPT} for clustering in ℓ_p under $(1 + \alpha, \epsilon)$ -approximation stability with $O(mk \log n)$ communication.*

Proof. The proof closely follows the previous proof, with necessary modifications. The properties in Lemma 9 are changed to the following.

- *Property 1:* For all x , for $\leq \frac{x\epsilon n}{\alpha^p}$ points v , $d(v, c_v)^p \geq \frac{\alpha^p \mathcal{OPT}^p}{x\epsilon n}$.
- *Property 2:* For $< 6\epsilon n$ points v , $d(v, c_i)^p - d(v, c_v)^p \leq \frac{\alpha^p \mathcal{OPT}^p}{2\epsilon n}$ for some $c_i \neq c_v$.

Then the proof for Lemma 9 remains unchanged. Now Algorithm 3 runs Algorithm 1 using a sequential β -approximation algorithm. When $p < \log n$, $\beta \in \Theta(1)$ [24]. Then Algorithm 3 sets $t = \left(\frac{\alpha^p \mathcal{OPT}^p}{2 \cdot 18^p \beta^p \epsilon n}\right)^{\frac{1}{p}} = \frac{1}{18} \left(\frac{\alpha^p \mathcal{OPT}^p}{2 \beta^p \epsilon n}\right)^{\frac{1}{p}}$. We define good points as $d(v, c_v) < t$ and $d(v, c_{v,2}) > 17t$, where $c_{v,2}$ denotes the second-closest optimal center to v . As before, B denotes the set of bad points.

Now we claim that $|B| < (6 + \frac{2 \cdot 18^p \beta^p}{\alpha^p})\epsilon n$. From the two properties, all but $(6 + \frac{2 \cdot 18^p \beta^p}{\alpha^p})\epsilon n$ points have $d(v, c_v)^p < \frac{\alpha^p \mathcal{OPT}^p}{2 \cdot 18^p \beta^p \epsilon n}$ and $d(v, c_{v,2})^p - d(v, c_v)^p > \frac{\alpha^p \mathcal{OPT}^p}{2 \beta^p \epsilon n}$ from which it follows that $d(v, c_v) > t$ and $d(v, c_{v,2}) > 17t$.

Now we have the same structure as in Theorem 7, where the good points are at distance t to their center, and at distance $17t$ to their next-closest center. The analysis carries through, and the error grows by a factor of $(\frac{18}{\alpha})^p$. \square

Theorem 8 (restated). *There exists an algorithm which outputs a clustering that is $O(\epsilon)$ -close to \mathcal{OPT} for k -median under $(1 + \alpha, \epsilon)$ -approximation stability with $O(m^2 kd + mk \log n)$ communication if each optimal cluster C_i has size $\Omega(\epsilon n(1 + \frac{1}{\alpha}))$.*

Algorithm 5 Distributed k -center clustering under 2-approximation stability

Input: Distributed points $V = V_1 \cup \dots \cup V_m$, optimal cost r^* .

1: For each machine t ,

Threshold using distance $2r^*$. Send one point per connected component to machine 1.

2: Given the set of points received, A , threshold using distance $2r^*$ again.

3: Pick one point per connected component, $X = \{x_1, \dots, x_k\}$.

Output: Centers $X = \{x_1, \dots, x_k\}$

Proof. The algorithm is as follows. First, run Algorithm 3. Then send X' to each machine i , incurring a communication cost of $O(m^2kd)$. For each machine i , for every point $v \in V_i$, calculate the median distance from v to each cluster $C(x_j)$ (using the weights). Assign v to the index j with the minimum median distance. Once every point undergoes this procedure, call the new clusters X_1, \dots, X_k , where X_j consists of all points assigned to index j . Now we will prove the clustering $\{X_1, \dots, X_k\}$ is $O(\epsilon)$ -close to the optimal clustering. Specifically, we will show that all are classified correctly except for the $6\epsilon n$ points in the bad case of Property 2 from Lemma 9.

Assume each cluster $C(x_j)$ contains a majority of points that are $2t$ to a point in H_j (we will prove this at the end). Given a point $v \in C_j$ such that $d(v, c_i) - d(v, c_j) > \frac{\alpha w_{avg}}{2\epsilon}$ for all $c_i \neq c_j$ (Property 2 from Lemma 9), and given a point $u \in C(x_j)$ that is at distance $2t$ to a point $u' \in H_j$, then $d(v, u) \leq d(v, c_j) + d(c_j, u') + d(u', u) \leq d(v, c_j) + 3t$. On the other hand, given $u \in C(x_{j'})$ that is at distance $2t$ to a point $u' \in H_{j'}$, then $d(v, u) \geq d(v, c_{j'}) - d(c_{j'}, u') - d(u', u) > 18t + d(v, c_j) - 3t \geq d(v, c_j) + 15t$. Then v 's median distance to $C(x_j)$ is $\leq d(v, c_j) + 3t$, and v 's median distance to any other cluster is $\geq d(v, c_j) + 15t$, so v will be assigned to the correct cluster.

Now we will prove each cluster $C(x_j)$ contains a majority of points that are $2t$ to a point in H_j . Assume for all j , $|C_j| > 16|B|$. It follows that for all j $|H_j| > 15|B|$. From the proof of Theorem 7, we know that $(\sum_j H_j \setminus (\sum_i C(v_j^i))) \leq 3|B|$, therefore, for all j , $H_j^A > 12|B|$, since H_j^A represents the points in A which are $2t$ to a point in H_j . Again from the proof of Theorem 7, the clustering $\{H_1^A, \dots, H_k^A\}$ is $9|B|$ -close to $X' = \{C(x_1), \dots, C(x_k)\}$. Then even if $C(x_j)$ is missing $9|B|$ good points, and contains $3|B|$ bad points, it will still have a majority of points that are within $2t$ of a point in H_j . This completes the proof. \square

B.1 k -center

In this section, we show a simple distributed algorithm to achieve the exact solution for k -center under $(2, 0)$ -approximation stability. Even in the single machine setting, it is NP-hard to find the exact solution to k -center under $(2 - \epsilon, 0)$ -approximation stability unless $NP = RP$ [9], therefore our result is tight with respect to the level of stability.

Given a 2-approximation stable k -center instance S , we assume our algorithm knows the value of \mathcal{OPT} . Since this value is some distance between two points in S , we can use binary search to efficiently guess this value; call it r^* .

Theorem 16. *Algorithm 5 outputs the optimal solution for k -center under 2-approximation stability, with $O(mkd)$ communication.*

Proof. First we prove the following: $u, v \in V$ are from the same optimal cluster iff $d(u, v) \leq 2r^*$. The forward direction is straightforward from the triangle inequality: if $u, v \in C_i$, then $d(u, v) \leq d(u, c_i) + d(c_i, v) \leq 2r^*$. For the reverse direction, assume there exist $i \neq j$, $u \in C_i$, $v \in C_j$ such that $d(u, v) \leq 2r^*$. Then consider the set of optimal centers $\{c_l\}_{l=1}^k$, but replacing c_i with u . Then for all

$u' \in C_i$, $d(u, u') \leq d(u, c_i) + d(c_i, u') \leq 2r^*$. Furthermore, $d(u, v) \leq 2r^*$. So the optimal clustering, but with v moved from C_j to C_i , achieves cost $2r^*$. This contradicts $(2, 0)$ -approximation stability.

Therefore, given any subset $V' \subseteq V$ of the point set, thresholding using distance $2r^*$ will exactly cluster the points into C'_1, \dots, C'_k such that $C'_i \subseteq C_i$ for all i . It follows that there is exactly one point per cluster in the outputted centers X . Since a point $u \in X \cap C_i$ is distance $2r^*$ from every point in C_i , X is a 2-approximation. Then by the definition of approximation stability, X defines the optimal clustering. \square